Perspectives in processing large amounts of information using Cloud

María A. Murazzo, Nelson R. Rodríguez, Daniela A. Villafañe, Facundo J. Gonzalez

Departamento e Instituto de Informática, UNSJ
[Department and Institute of Computing]
San Juan, C.P. 5400, Argentina
UNSJ, Cereceto y Meglioli, C.P. 5400
Rivadavia, San Juan, Argentina

ABSTRACT

Nowadays it is easy for the users who live in a digital era, to access social networks not only from their computers, but also from their mobile phones, to upload pictures instantaneously, to send messages on Whatsapp or to share their location with other users.

At the same time, the business world is also changing at high-speed. Dematerialization has already affected different business sectors; it has even put an end to a model that has been used for years.

Cloud Computing allows technology to access in the economy and society, not only allowing users to be connected to this new digital world through their mobile devices, but also to do it through any kind of device, which is commonly called Internet of Things. This will create a large amount of digital information which will need the storage and processing capacity of large volumes known as Big Data.

The aim of this work is to determine the ideas of investigation to allow a proper analysis of large volumes of data which has been created to improve the efficiency and effectiveness of the decision making system.

Key Words: Cloud Computing, Big Data, NoSQL

1. INTRODUCTION

TIC (Information and Communications Technology) has transformed the ways of reading and writing, creating a system where social, cultural and educational activities are connected in their development.

Nowadays, the production, distribution and consumption systems have a digital nature. Markets and companies are experiencing a technology and social based transformation which main result is the growth of data inside and outside of business systems.

GPS localization of mobile devices, likes on Facebook, e-commerce transactions, videos surveillance camera, instant messaging, etc. are an example of the challenge set by technology and the capacity to make use of it. An important matter is that 90% of the available information has been generated over the last two years, and it is mostly digital information.

Two terabytes of new data is generated everyday, 10% of that information is structured, bibliographically controlled; the remaining 90% is not structured at all.

The growth of the information available with each new media that has come to existence has lead to a great growth of the information available.

According to Gartner (www. gartner.com), in 2015 almost 3 trillion people will be online producing around 8 zettabytes of information.

According to IBM, human beings have created sine their beginning from their history to 2003, around 5 exabytes of information, 5 billions of gigabytes. In 2011 the same amount of information has been generated every two days and in 2013 that amount of information will be produced every ten minutes.

According to IDC (www. idc.com), the volume of digital data will reach 35 Zbytes, 44 times more than in 2009. This growth is characterized for being formed by an 80% of unstructured data.

Organizations face the challenge of capturing, transforming, analysing and storing non-traditional and varied information with traditional systems that are not able to make a proper use of them in time and manner.

1.1 What is Cloud Computing?

In recent years, technologies have evolved for the markets according to TIC, such as telephone services, telecommunications, data center, etc.

Why shouldn't I connect to internet and have someone running all the computer services that the company needs in an easy way and charge me monthly for that? In that way, all computing elements will become another utility.

This is not a new idea; this concept has been in progress over the last few years, since it is the convergence of revolutionary models such as Utility Computing, On Demand Computing, Elastic Computing or grid computing [1].

Internet is generally seen and considered as a big cloud where everything is connected and when connected to it all the services required are available. This system is known as Cloud Computing, which is similar to the systems that have been mentioned before, but this one is improved by the virtualization technology [2].

> The main characteristic of the Cloud Computing concept is the transformation of the traditional ways of how companies use and acquire new sources of information technology (IT).

Cloud Computing represents a new kind of value to the network computing. It provides more efficiency, massive scalability, faster and easy software development. The new programming models and the new IT infrastructure will lead to new business models.

Cloud Computing is a provider of IT resources which improves the provision of IT and business services, making it easier for the user final user and for the provider of the service.

One of the advantages of the companies which decide to incorporate to their activities services provided by Internet is the possibility of reducing the cost of specialists in technology, installation, software and most important of all, maintenance activities; for seeing the benefits of this investment is immediate, since there is no need for pre-installation or configuration at all.

This is processed in a reliable and secure way, with elastic scalability that is capable of accepting the unexpected changes on the demand, without a higher cost on the administration.

The basic characteristic of this model is that computing resources and services, such as infrastructure, platform and applications, are offered and consumed by services on Internet and users do not need to know what happens backstage.

Cloud Computing is service such as aaS or as a Service which sometimes it is referred as XaaS or EaaS to mean everything as a service [3].

This new phenomenon coincides in time with the massive diffusion of mobiles, making possible a future with extreme mobility in which the whole world is connected (without cloud, it wouldn't be possible to store and transfer the information of the 900 million of Facebook users, and many of them are connected using their mobiles). We have to consider that since 2010 more mobiles are manufactured than computers, computers have been replaced by devices that allow the user to be connected anywhere [4].

These devices connected to the cloud, will create a new relationship between the users and their devices, and to improve the quality of that interaction, the information will be used for context and preferences of the user, such as location services, augmented reality in mobile devices and mobile e-commerce. Companies will be able to anticipate to the needs of the user and offer services and products more appropriate and personalized (for instance, marketing and commercialization based on geolocalization to find offers on real time or using promotional codes, or discounts, etc).

In that sense, cloud computing can connect not only millions of users who use it through their computer or mobile device, but also with objects with capacity for communication at any time and anywhere. This is called Internet of Things, which implies that every object can be a source of data and everything can be monitored in time and space. These are low cost chips embedded in almost any object of consumption (such as refrigerator, coffee machines, cars, etc) [6].

It can be said that Cloud Computing has also made that smart phones and other devices of massive consumption replace the computer in the digital era. Cloud Computing is creating a new era which gives the user more flexibility for the devices to be able to do everyday activities.

Besides, the new ways of interacting, such as the ways provided by tactile technology as well as the voice recognition and contextual knowledge, offer a good interaction between the user and their devices, and a huge volume of varied information is stored in the cloud.

The intelligent use of this information will be able to transform society, which in the business sector; will provide the creation of new models, such as an improvement of the existing ones, whit lower costs and lesser risks.

2. LARGE AMOUNT OF INFORMATION

To understand how easy is to create information, it is necessary to analyse the social media, for example Twitter, which has more than 90 millions of tweets per day, which represents a total of 8 terabytes of information everyday. The information of web transactions has increased significantly, nowadays, Wal-Mart, the largest retailer in the world, administrates one million of transactions by hour which increases a data base valued on 2.5 petabytes. An example in the scientific area is the collider of particles CERN, which can eve create 40 terabytes of information per second during the experiments. In networking, the information registered by a system of research and network management can reach terabytes in two days.

This creates a great growth of the information available and leads to the next big tendency of ICT, known as Big Data. According to some studies, in 2015 there will be more than 7.910 exabytes of information in the planet, (in 2005 there was only 130 exabytes). To generate exabytes of information is really easy but to process them and transform them in valued data is more complex. The issue is not to find information but to know what to do with it.

As it happened with Cloud Computing, there is not only one definition of agreement for Big Data. According to IDC, Big Data is a new generation of technologies and architectures designed to get economical value of large amount of varied information making able a new capture, identification and fast analysis. Big Data is characterized for having four dimensions:

- *Volume:* It makes reference to the need of intense and complex processing of subset of data that contains a large amount of valued information for an organization through technologies of Big Data.
- *Variety:* due to the growth of the channels of interaction with costumers, employees, providers and business units, valued information is a result of a combination of data with different origins and typology that may be structured, semi-structured or unstructured.
- *Velocity:* even though the business cycles have accelerated, not all the information of an organization is key to understand the urgency of the analysis. To understand at which speed it is necessary to work (from mass processing to the transference of data) is associated to the requirements of the processes and users.
- *Value:* in Big Data the value makes reference to the benefits that result from the use of Big Data (reduction of costs, operational efficiency, improvements in businesses).

When analysing this technology, two points have to be considered: the way to find value en every data (data science) and how to manage them. It is necessary to take into account that even though there is enough storage capacity, the time to access data may be long, and the time needed to transfer data is as important as the storage capacity [6].

2.1Data storage and analysis

Until a few years ago, the data base theme has been determined by what is known as SQL10 (Structured Query Language). This data base, which is commonly known as relationships, comply with the ACID rules (Atomicity, Consistency, Isolation and Durability) what allow them to ensure that the information is stored properly and related to a structure based on tables which have rows and columns.

But Big Data also has a problem; data base can not manage the size, complexity and variety of formats and the velocity of data delivery that requires some applications, such as apps on line with thousands of users and millions of enquiries per day. The current storage systems have two restrictions:

1) On one hand, these systems are not able to comply with certain requirements, such as offering a low time of answer, which some applications are able to give. In such cases, data is being stored in memory (known as in-memory). This kind of applications will be benefited from the improvement seek in storage.

2) On the other hand, some applications require a large amount of information that cannot be stored and processed using the traditional data base.

For this reasons, new data base has been created, known as NoSQL, which are able to solve the issues of scalability and performance that Big Data presents. NoSQL brings together the different solutions that cell-centered database for not being relational, distributed and scalable in horizontal.

NoSQL database does not mean SQL is not necessary, but there are better solutions for certain problems and applications. Due to this issue, NoSQL can be seen as not only SQL. The idea is that for certain cases, there is a need to identify many of the limits of the conventional database and the way of accessing to them. For example, one may think about large documents with defined fields not well determined that may even change with the passing of time, instead of having tables with set rows and columns.

Actually, NoSQL database is not only a database, it is a storage system presented to manage data that is structured that may be flexible, such as Picasa albums, videos stored in YouTube, files of big systems, captures of network transit, etc [7].

The information provided by Big Data helps organizations to know how users and market act, the use of these large volumes of data can give support marketing campaigns and strategies to facilitate the procedures of quality control, help in revision, improve customer service and comply with the rules, manage the risks, etc, in short, companies are able to get better competitive advantages.

To process large amounts of information well distributed and unstructured presents a big challenge, and in these cases the traditional methods of processing are not useful either. In such cases, it is necessary to have the ability of transforming data in information, and information in knowledge, in order to optimize the business procedures with such knowledge.

To be able to process large amount of information, Google created MapReduce, a program which processes the data of Google. But the implementation of Hadoop MapReduce, by Yahoo, has provided with open-source tools for processing large amounts of information that the majority of other systems of other companies are using.

3. BIG DATA TECHNOLOGIES

In recent years, many algorithms have been implemented to process large amounts of information. Many of these calculations are simple. However, input data is generally large, and the calculations have to be distributed through many machines in order to be able to do it in a short period of time. The fact of how to make the calculation simultaneously, distribute the data and manage the errors, make the original algorithm more difficult by introducing a large amount of complex codes to address those issues.

To solve this issue, a new abstraction has been thought in order to express simple calculations that are trying to make, but hiding the complex details of parallelism, tolerance to errors, data distribution and balance of load in a library. MapReduce would be able to solve these problems.

MapReduce is the name given to the combination of two separate procedures necessary to get the value of the origin of different information. Map procedure works as filter and gives value to certain keys for only one document. Reduce procedure is in charge of storing and combining the keys of several documents to create an only one value for each key from the several generated values. The good thing about this way of working is that it is easy to implement on a cluster.

MapReduce is a framework introduced by Google in 2004 to provide support to parallel computing on large amounts of information in many devices. The process may be done in stored data, such as in a file system (unstructured) or inside a database (structured).

One advantage of MapReduce is that it allows the distributed processing of map and reduce functions As long as each assignment operation is independent, all map operations can be processed simultaneously, though in practice this parallelism is limited by the source of the information and the CPU’s number. In the same way, some reduce operations can be processed in the Reduce phase. Its requirements are that the output of Map that share the same key is presented in the same reduce operation at the same time.

In this way, many servers may use MapReduce to order a petabyte of data in few hours. This method also provides the possibility to recover of an error, the work may be reprogrammed, considering the input data is still available [8].

Fortunately, Hadoop is inspired in the project of Google File System (GFS) and in the MapReduce programming, which consist of dividing two tasks (mapper – reducer) to manage the distributed data having a high parallelism in the procedure. Hadoop is a project presented by Apache Software Foundation which objective is the development of parallel processing applications that uses MapReduce and allows applications to work with thousands of nodes and petabytes of information [9].

One of the problems that MapReduce presents is the incompatibility en the way in which data is processed in SQL data bases. Due to this problem, different projects have been presented, such as Hive, which is a data warehouse based on Hadoop which was developed by Facebook and now is an open-source project inside Hadoop. The main idea of Hive is to offer the possibility to the users to write their issues in SQL, that later are converted into MapReduce in a clear way for the programmer. This allows SQL programmers who have no experience in MapRedcue to use it and include it in their regular activities [10].

Pig Latin is also a package in Hadoop that was originally developed by Yahoo that offers high level language to describe and execute works of MapReduce. Its objective is to make Hadoop accessible for the developers used to the SQL data management, which provides two interfaces: one interactive and the other one able to use it in Java [11].

3.1Working with BigData besides MapReduce

There are some applications where MapReduce is not a good option (such as Percolator, Dremel and Pregel). One of the solutions is to make MapReduce faster on moving the data in the memory but there are certain tasks where its structure makes it difficult for MapReduce to scale. MapReduce have two serious issues: a high latency and wearing out (it has to recalculate all the information even when only one small fraction of it has changed).

Percolator is a system for the processing of changes in big quantities. By the substitution of an indexation system based on sets with one of processing such as Percolator, it can make the process faster and reduce the time of analysis. The best candidates for this one are great indexes, where the factor of performance can be 100%. The great advantage of Percolator is that time of indexation is proportional to the size of the page, not to the size of index [12].

Dremel is for the ad-hoc analysis. It is a scalable system, interactive ad-hoc consultation for the analysis of reading nested data. With the combination of multi-level execution trees and the distribution of column-oriented data, it is able to execute consultations of adding tables of a trillion of rows in just seconds. Dremel is a 100 times faster than MapReduce. Its architecture is similar to the one of Pig and Hive, but instead of MapReduce, its motor is based on aggregation tree [13].

Pregel is a system for processing and analysing large-scale graphs. It is designed to execute fast graph algorithms and has an API easy to use. It has architecture for an efficient implementation, scalable and tolerant to errors in a cluster of thousands of commodity machine. Graphs are everywhere, social networks, network topology, football matches, scientific articles, and the most important of all, the web. Pregel is a scalable infrastructure to exploit a large variety of graphs and programmes which are expressed as a sequence of iterations.

Many open-source projects have been based on the ideas and papers of Google. For example, ApacheDrill is a reimplementation of Dremelframework, also projects like Apache Giraph and GPS of Stanford are inspired in Pregel.

4. CONCLUSIONS

The size, the complexity of the formats and the velocity of delivery are grater than the capacity of the technologies of management of traditional data. So the use of new technologies is required to allow the management of large amounts of information, in this sense, new technologies have been developed that are able to execute a great impact, such as data base programmers in memory to address the new requirements of the information.

According to what has been considered, the need of integrating Cloud Computing and Big Data in order to be able to analyse large amounts of varied information and being able to make the proper analysis to get the required information in order to improve the decision-making process.

In that sense, the objective is to design the architecture necessary to provide the execution of algorithms for the simultaneous process and distributed as the analysis of massive groups of unstructured data. The resulting design will allow the analysis of scalability and performance, in distributed environments.

5. REFERENCES

[1] Marston, Li, Bandyopadhyay, Zhang, Ghalsasi. “Cloud computing — The business perspective”. Decision Support Systems 51 (2011) 176-180. Elsevier. 2011.

[2] Lu, Hai-shan, Ting-ting.”Research on Hadoop Cloud Computing Model and its Applications”.2012. Third International Conference on Networking and Distributed Computing

[3] Murazzo, Rodriguez, Segura, Villafañe. “Desarrollo de aplicaciones para Cloud Computing”. CACIC 2010. Morón. Oct. 2010.

[4] Murazzo, Rodriguez. “Mobile cloud computing”. WICC 2010. Calafate, Santa Cruz.

[5] Atzori, Iera, Morabito. The Internet of Things: A survey”. Computer Networks, Volume 54, Issue 15, 28 October 2010, Pages 2787–2805.

[6] Agrawal, Das, Abbadi. “Big data and cloud computing: current state and future opportunities”. 14th International Conference on Extending Database Technology. Pages 530-533 ACM New York, USA 2011.

[7] Martin, Chávez, Rodríguez Murazzo, Villafañe, Valenzuela. “Bases de Datos NoSql en Cloud Computing”. WICC 2013. Paraná, Entre Ríos.

[8] Dean, Ghemawat. “MapReduce: Simplified Data Processing on Large Clusters”. http://research.google.com/archive/mapreduce.html.

[9] Apache Software Foundation. ”Apache Hadoop”. http://hadoop.apache.org

[10] Apache Software Foundation. ”Apache Hive”. http://hive.apache.org

[11] Apache Software Foundation. ”Apache Pig”. http://pig.apache.org

[12] Peng, Dabek. “Large-scale Incremental Processing Using Distributed Transactions and Notifications”. Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation, USENIX (2010). http://research.google.com/pubs/pub36726.html.

[13] Melnik, Gubarev, Jing Long, Romer, Shivakumar, Tolton, Vassilakis. “ Dremel: Interactive Analysis of Web-Scale Datasets”. Proc. of the 36th Int'l Conf on Very Large Data Bases (2010), pp. 330-339. http://research.google.com/pubs/pub36632.html.

[14] Austern, Bik, Dehnert, Horn, Leiser, Czajkowski. “Pregel: A System for Large-Scale Graph Processing”. 28th ACM Symposium on Principles of Distributed Computing (2009), pp. 6-6. http://research.google.com/pubs/author38459.html.